Reference: *Biol. Bull.* **172:** 269–283. (June, 1987)

THE EVOLUTION OF THE VERTEBRATE PLASMA PROTEINS*

RUSSELL F. DOOLITTLE

Department of Chemistry, University of California, San Diego, La Jolla, California 92093

ABSTRACT

The appearance of vertebrate animals some 450 million years ago was heralded by the invention of a number of new proteins that are found in the blood plasma, including albumin and other transport proteins, the non-enzyme blood clotting and complement proteins, and a host of protease inhibitors. Comparisons of present day amino acid sequences from various species allow us to look back to how and when many of these proteins originated.

INTRODUCTION

It is particularly gratifying for me to present this Friday Evening Lecture on the evolution of vertebrate plasma proteins, since it is a project I began many years ago here at the MBL. At the time I was a graduate student at Harvard in a laboratory devoted to the study of human blood proteins, and, after having spent a summer in Woods Hole working in the area of comparative physiology, it was only natural that I should consider a comparative study of blood proteins from other vertebrates. The question immediately presented itself as to where, in the biological world, the characteristic proteins of blood are found, and from this gradually emerged the more profound questions as to when and how they evolved.

A typical mammalian blood plasma—the fluid in which blood cells are suspended—contains upwards of 600 protein components, as can be shown by high resolution two-dimensional electrophoresis (Anderson *et al.,* 1984). Some of these materials are doubtless only altered forms of particular proteins that result from various processing events; still, the multitude of individual types is impressive. In mammals, a single protein, albumin, accounts for half the protein mass of the plasma, however, and five more—haptoglobin, fibrinogen, transferrin, α-1-antitrypsin and α-2-macroglobulin—make up another quarter; lipoproteins and immunoglobulins constitute another large fraction (Table I). This evening I am going to concentrate on the origins of some of these most abundant proteins, although I will touch upon a few of the less abundant ones in passing.

Regarding the phyletic distribution of the plasma proteins, we can state flatly that many of these proteins, but certainly not all of them, are found in the blood plasmas of all the major classes of vertebrates, from fish to mammals. It has been alleged by some that albumin is absent from the blood of fish, although some workers have reported that it is present in assorted teleosts in small amounts (for a review, see Doolittle, 1984). There is general agreement that amphibians, reptiles, and birds all have albumins that are similar to the mammalian type. Only a few of the vertebrate plasma proteins have counterparts among the invertebrates, and, as we shall see, albu-

Received 3 March 1987; accepted 23 March 1987.

* This article is reproduced from the Friday Evening Lecture delivered by Dr. Doolittle at the Marine Biological Laboratory on 8 August 1986.

TABLE I

Most abundant proteins found in mammalian blood plasma

Protein	Abundance* (gm/liter)	Internal duplications	Carbohydrate
Albumin	45	Yes	No
Immunoglobulins	15	Yes	Yes
Lipoproteins	10	Yes	Yes
Haptoglobin	6	Yes	Yes
Fibrinogen	3.5	Yes	Yes
Transferrin	3	Yes	Yes
α_1-Antitrypsin	3	No	Yes
α_2-Macroglobulin	2.5	No	Yes
	87.5		

* These eight (sets of) proteins account for more than 95% of the mass of plasma proteins in mammals. The values given are approximate and vary somewhat from species to species and among individual species.

min is not among them. We'll return to the matter of occurrence shortly, but there are a few definitional matters to be settled first.

FUNCTIONS OF THE PLASMA PROTEINS

The plasma proteins can be grouped functionally. Quite apart from albumin, there are numerous transport proteins, including transferrin (iron transport) and the lipoproteins. Indeed, any potential metabolite that is poorly soluble in plasma is likely to have a corresponding protein for rendering it soluble so that it can be transported from one place to another.

In addition to the transport proteins, the plasma contains numerous polypeptide hormones and growth factors, many in precursor forms awaiting proteolytic activation upon the appropriate signal. There are also numerous defense proteins, including the immunoglobulins and the associated complement system, and there is a set of proteins whose function is to prevent loss of the blood: the coagulation proteins. The latter is a special interest of mine, and I will be giving it an undue amount of attention this evening.

The complement and blood coagulation systems, and some other schemes in the plasma, depend upon a complex plan of proteolytic activation for their normal operation. Indeed, much of the extracellular way of life in multicellular creatures is regulated by a judicious and limited proteolysis. Whereas intracellular regulation depends largely on ATP-dependent phosphorylation (kinases) and subsequent hydrolysis (phosphatases), extracellular regulation derives its energy potential at the time of protein biosynthesis. It is then "leaked out" at some later time by a gradual and successive proteolysis.

As a consequence, blood plasma is rich with proteases and protease precursors. There are very many related serine proteases involved, and some sulfhydryl proteases. In order to keep the system stable, there are large amounts of general and specific protease inhibitors. Two of these, α-2-macroglobulin and α-1-antitrypsin, appear on our list of the most abundant plasma proteins (Table I), and they will figure heavily in our evolutionary discussion.

HOW NEW PROTEINS ARE FORMED

At this point let me make some sweeping statements about the evolution of proteins in general. First, most new proteins evolve from old proteins as a result of gene

TABLE II

*Divergence times of some major vertebrate groupings**

Vertebrate groupings	Millions of years ago
Old World monkeys/humans	28 ± 2
Artiodactyls/primates	80 ± 10
Primates/rodents	80 ± 10
Birds/mammals	200 ± 20
Reptiles/mammals	200 ± 20
Amphibia/mammals	320 ± 30
Bony fish/mammals	375 ± 35
Cyclostomes/mammals	450 ± 50

* Groupings are based on the fossil record.

duplication. DNA has a propensity for duplicating itself under all sorts of conditions, and segments of all sizes are constantly being randomly and tandemly repeated. There are many known mechanisms for this excessive internal duplication, including unequal and homologous crossing over and other genetic delinquencies, but the important thing for the present discussion is that these duplications are observed in all living forms. If the region of duplicated DNA is short and inside the boundaries of the gene for a particular protein, then the protein is lengthened by the process, some portion of it now showing up twice or more. If the DNA duplication encompasses the entire gene—starts and stops included—then for a brief moment two genes will exist where there used to be one, and two separate products may be issued. Only one of these need be subject to natural selection, of course, and the other will be free to change, mostly as a result of errant base substitutions leading to single amino acid replacements. Usually such duplicons are destined for the genomic scrap-pile, since sooner or later a mutation will occur that will decommission the whole enterprise, but occasionally some new function—with benefit to the organism—will be encountered by chance. And with that, a *bona fide* "new" protein is eligible for the registry. It is related to the "old protein" of course, and depending on the rate of amino acid replacement along their divergent courses, the two will, for a long time, be recognizable merely on the basis of their amino acid sequences alone. By "a long time" I am implying time frames of a 100 million to several billion years (Table II).

SETTING MOLECULAR CLOCKS

There are two independent means at our disposal for finding when a "new protein" made its appearance during evolution. We have already alluded to one of these, which simply amounts to surveying the biological community to see which creatures have the protein in question. If all mammals have some protein—*e.g.,* lactalbumin, which is found in milk—and none of the other classes of vertebrates have this protein, then we can presume that lactalbumin was "invented" some time around or just before the appearance of mammals, about 200 million years ago. If all vertebrates except fish have some particular protein, then we would expect that protein must have evolved about 375 million years ago, since that is when fish and the other vertebrates last shared a common ancestor (Table II).

The second way to gauge when a gene duplication leading to a new protein occurred, involves the comparison of amino acid sequences and the use of the "molecular clock." The method can also be used to find when an internal duplication leading

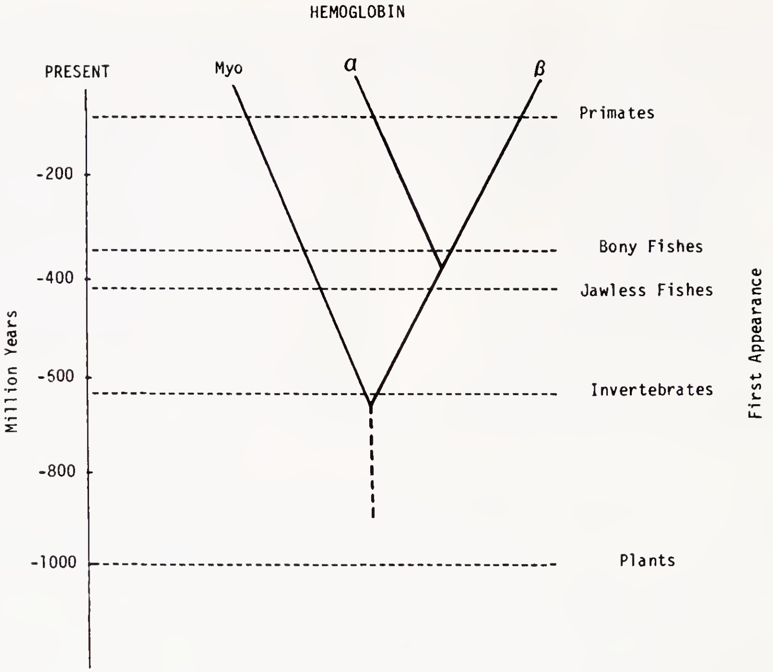

FIGURE 1. Gene duplication event that led to the existence of hemoglobin α and β chains. The timescale is based on how fast the amino acid sequences are changing as determined by species comparisons (from Doolittle, 1984).

to an internal repeat may have taken place. Although this is a very useful method, it is often subject to a certain amount of misinterpretation, and it is worth our while to remind ourselves how it works. First, it is a fact that the sequences of various proteins change at different but characteristic rates. For many proteins, as we shall see, the rate of change is really quite constant. In some situations, nonetheless, the rate of change of a particular type of protein may speed up or slow down, and we must be on the watch for a certain amount of quirkiness in molecular clocks. When they run smoothly they'll be in accord with what we find by the "occurrence method."

Consider a well known example. The α and β chains of vertebrate hemoglobin have been sequenced from many different species in all five classes. By consulting the divergence times in Table II, which are based on the fossil record, and quantitatively comparing the sequences from members of each group, we can estimate that the two chains are each changing at a rate of about 11 amino acid replacements per 100 residues per 100 million years. Today the two sequences in most vertebrates are 55% different (45% identical). As it happens, sequence change follows an exponential course (because of back mutations and multiple changes at the same site), and a 55% difference is actually equivalent to 90 actual changes for every 100 residues. At a rate of 11 per 100 million years, then, the time since the duplication that gave rise to divergence must be about 400 million years (keep in mind both proteins are changing). That being the case, we can make a prediction, assuming the clock has been ticking accurately. Any vertebrates that diverged from the mainline longer than 400 million years ago ought not to have the benefit of that duplication (Fig. 1). Indeed, cyclostomes (lampreys and hagfish) branched off about 450 million years ago, and those creatures have single-chained hemoglobins: no β chains!

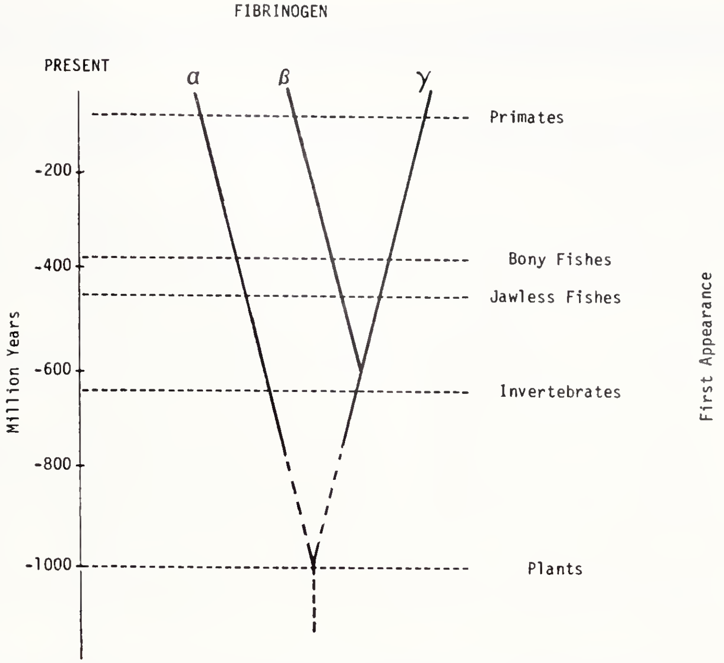

FIGURE 2. Gene duplication events that lead to three different polypeptide chains in vertebrate fibrinogens. See also Figure 1 (from Doolittle. 1984).

FIBRINOGEN

I began my studies on the evolution of the plasma proteins with a consideration of the proteins involved in blood coagulation, a phenomenon which in mammals involves the interplay of a dozen different protein factors and that culminates in the conversion of the soluble protein fibrinogen into the insoluble gel fibrin. How could this system evolve? Of what utility would any portion of the cascade be without the remainder? These were the original questions, and I must admit at once that they remain mostly unanswered in a strict sense, although at this point a number of feasible scenarios can be drawn.

Blood clotting follows a similar pattern in all vertebrates, from the cyclostomes to the mammals. Lampreys have a fibrinogen molecule that is fundamentally the same as the human kind. It has three polypeptide chains (α, β, and γ), and it is clotted by the proteolytic enzyme thrombin, which in lampreys is also similar to its human counterpart (Doolittle *et al.,* 1962). During the 1970's the complete amino acid sequence of human fibrinogen was unraveled, and the results confirmed earlier speculation that the three polypeptide chains were descended from a common ancestral type (Fig. 2). Thus, the β and γ chains are about 35% identical, and both of the latter are recognizably similar to parts of the α chain, although the resemblance is somewhat lower in the latter cases. Apparently the duplication leading to the divergence of the α and non-α chains occurred much longer ago than the one leading to β and γ chains.

Recently we completed the sequences of the β and γ chains of lamprey fibrinogen; in each case the sequences are just about 50% identical with the corresponding human sequence (Figs. 3, 4). That these two independent gene products have experienced such similar amounts of change proves that the molecular clock in this particular

```
                  .  .  .. ....... .. .. .   .. .       . .   .   .
Human     YVATRDNccILDERFGSYcPTTcGIADFLSTYQTKVDKDLQSLEDILHQVENKTSEVKQL
Lamprey    QVRDLKQcSNDPEFGRYcPTTcGVADVLSKYAKGVDEDSSFIDSVLTQLAAKHGIVEGN

                             .       . ...    .     ..  .  . .    . .
Human     IKAIQLTYNPDESSKPNMIDAATLKSRKMLEEIMKYEASILTHDSSIRYLQEIYNSNNQK
Lamprey   VNIVNEDVRITRDEAQIIKDSGQKTVQKILEEVRILEQIGVSHDAQIQELSEMWRVNQQF

             .          .  .. ..    .  .......     . . .... ...... . .
Human     IVNLKEKVAQLEAQcQEPcKDTV  QIHDITGKDcQDIANKGAKQSGLYFIKPLKANQQF
Lamprey   VTRLQQQLVDIRQTcSRPcQDTTANKISPITGKDcQQVVDNGGKDSGLYYIKPLKAKQPF

          .. ...   ...... . . .... .   .  . .... . .. .............
Human     LVYcEIDGSGNGWTVFQKRLDGSVDFKKNWIQYKEGFGHLSPTGTTEFWLGNEKIHLIST
Lamprey   LVFcEI ENGNGWTVIQHRHDGSVNFTRDWVSYREGFGYLAPTLTTEFWLGNEKIHLLTG

          . .  . ..  . ..      ...  ..  .. . ... .     .... ..........
Human     QSAIPYALRVELEDWNGRTSTADYAMFKVGPEADKYRLTYAYFAGGDAGDAFDGFDFGDD
Lamprey   QQA  YRLRIDLTDWENTHRYADYGHFKLTPESDEYRLFYSMYLDGDAGNAFDGFDFGDD

          . ... . . .. ...   .... .. ............ ........ .. .. . .
Human     PSDKFFTSHNGMQFSTWDNDNDKFEGNcAEQDGSGWWMNKcHAGHLNGVYYQGGTYSKAS
Lamprey   PQDKFYTTHLGMLFSTPERDNDKYEGScAEQDGSGWWMNRcHAGHLNGKYYFGGNYRKTD

              .. .......  .... . ....  .  .   . .↓.    .
Human     TPNGYDNGIIWATWKTRWYSMKKTTMKIIPFNRLTIGEGQQHHLGGAK̄QAGDV
Lamprey   VEFPYDDGIIWATWHDRWYSLKMTTMKLLPMGRDLSGHGGQQQSK̲GNSRGDN
                                                   ↑
```

FIGURE 3. Alignment of lamprey and human fibrinogen γ-chain sequences. There are 205 identities among the 408 aligned residues, which amounts to 50.2% identity (from Strong *et al.*, 1985).

molecule is running smoothly. The data also indicate that the gene duplication which led to β and γ chains must have occurred about 600 million years ago. The fact that the α chains are more divergent implies that the first duplication, leading to α and non-α chains, may have occurred as much as a billion years ago, although in this case we must be more cautious, since there are portions of fibrinogen α chains that are changing significantly faster than the rest of the molecule, and the elapsed time involved may therefore be significantly less.

In either case, there is a paradox here, for although the sequence data indicate that fibrinogen was evolved somewhere between 600 million and a billion years ago, nobody has yet observed such a molecule among the invertebrates or protochordates, and it is not for a lack of looking. It is true that some arthropods have a protein in their hemolymph that can be gelled under appropriate circumstances, but it is a fundamentally different molecule and does not share ancestry with vertebrate fibrinogen (Fuller and Doolittle, 1971; Doolittle and Fuller, 1972). So, more than 25 years after first asking when and where did the vertebrate fibrinogen molecule evolve, I am still begging the question. The tools for searching are much improved now, however, and we are now using recombinant DNA techniques to probe the genomes of various protochordates and echinoderms. The advantage is that even if the ancestral protein is made in amounts too small to have been detected by ordinary means, or in an intracellular setting where we wouldn't have seen it, we should be able to find its genes with appropriate DNA probes.

```
          . ...    .   . ..        .    .   .            . ...     ...
Human     GHRPLDK KREEAPSLRPAPPPISGGGYRARPAKAAATQKKVERKAPDAGGcLHADPDLG
Lamprey   GVRPLPSGTRVRRPPLR  HRRLAPGAVMSRDPPASPRPQEAQKAIRDEGGcMLPESDLG

          ....... . . .. .  . .     .  .                 ..   .
Human     VLcPTGcQLQEALLQQERPIRNSVDELNNNVEAVSQTSSSSFQYMYLLKDLWQKRQKQVK
Lamprey   VLcPTGcELREELLKQRDPVRYKISMLKQNLTYFINSFDRMASDSNTLKQNVQTLRRRLN

                     . .           .     ..   . .. .. . .  .     . . .
Human     DNENVVNEYSSELEKHQLYIDETVNSNIPTNLRVLRSILENLRSKIQKLESDVSAQMEYc
Lamprey   SRSSTHVNAQKEIENRYKEVKIRIESTVAGSLRSMKSVLEHLRAKMQRMEEAIKTQKELc

            .... .  .....  .. . . .. ... . ....    .. . ...    ..... .
Human     RTPcTVScNIPVVSGKEcEEIIRKGGETSEMYLIQPDSSVKPYRVYcDMNTENGGWTVIQ
Lamprey   SAPcTVNcRVPVVSGMHcEDIYRNGGRTSEAYYIQPDLFSEPYKVFcDMESHGGGWTVVQ

          .. ...  . . .  ..  ... .    ..  .  .......     ...       .
Human     NRQDGSVDFGRKWDPYKQGFGNVATNTDGKNYcGLPGEYWLGNDKISQLTRMGPTELLIE
Lamprey   NRVDGSSNFARDWNTYKAEFGNIA FGNGKSIcNIPGEYWLGTKTVHQLTKQHTQQVLFD

          . .. .  . . .  .   ...  .   .  . . .....  .. .... ..........
Human     MEDWKGDKVKAHYGGFTVQNEANKYQISVNKYRGTAGNALMDGASQLMGENRTMTIHNGM
Lamprey   MSDWEGSSVYAQYASFRPENEAQGYRLWVEDYSGNAGNALLEGATQLMGDNRTMTIHNGM

           ... ..... .   .. . .. .. ..................... ..   .  ....
Human     FFSTYDRDNDGWLTSDPRKQcSKEDGGGWWYNRcHAANPNGRYYWGGQYTWDMAKHGTDD
Lamprey   QFSTFDRDNDNWNPGDPTKHcSREDAGGWWYNRcHAANPNGRYYWGGIYTKEQADYGTDD

          ............... . ..  ..  .
Human     GVVWMNWKGSWYSMRKMSMKIRPFFPQQ
Lamprey   GVVWMNWKGSWYSMRQMAMKLRPKWP
```

FIGURE 4. Alignment of lamprey and human fibrin β-chain sequences. There are 218 identities among the 443 aligned residues, which amounts to 49.2% identity (from Bohonus *et al.*, 1986).

ALBUMIN

As I mentioned earlier, not everyone in the comparative plasma protein field is convinced that fish have an albumin in their blood plasma, and it is often viewed as a protein that evolved as a part of the land invasion by vertebrates (again, for a review of opinions, see Doolittle, 1984). Its great abundance was long considered by physiologists as being primarily of osmotic importance ("Starling's Law of Colloid Osmotic Pressure"), even though many investigators recognized its important contributions to the transport of fatty acids and other nonpolar materials.

In the mid-1970's, Jim Brown (1976) reported the amino acid sequence of the bovine albumin molecule, and the data revealed that the protein had experienced a number of internal duplications. In particular, the sequence could be divided into three equivalent regions of about 190 amino acids each. The sequences of two of these were more similar then either was to the third, suggesting that there had been a doubling at one point, from a molecule of about 190 amino acids to one of 380, and then a second, incomplete, duplication that gave rise to the existing 580-residue structure. There was also evidence that the fundamental macrodomain structure (*ca.* 190 residues) had itself evolved as the result of internal duplications from a more primitive sequence of about 77 residues. Comparative studies on the human protein indicated that albumin is changing almost twice as fast as the hemoglobin polypeptides. Brown (1976) interpreted his data as indicating that the tandem duplications

TABLE III

*Comparison of lamprey and mammalian plasma albumins**

	Mammalian	Lamprey
Molecular weight	69,000	175,000
Water-soluble	Yes	Yes
Bind bromphenyl blue	Yes	Yes
Free-SH	1	2
Tryptophans	2–3	3–4
Disulfides	Rich	Rich

* From Kuyas *et al.* (1983).

leading to the three macrodomains occurred about 700 million years ago, a number that may be a little high, but one that obviously suggests that invertebrates ought to have albumin. Again, no one has ever observed a protein among the invertebrates that bears any resemblance to a vertebrate albumin.

My own calculation of when the albumin internal duplications ought to have occurred resulted in a somewhat more recent time than Brown had reckoned, and, accordingly, I considered the possibility that fish, and the lamprey in particular, might have a smaller, more ancient, molecule, corresponding to the one- or two-macrodomain stage. Imagine our surprise, then, when we found that one of the most abundant proteins in lamprey plasma has all the properties of a mammalian albumin (Table III) except that its molecular weight is 175,000, fully two and-a-half times *larger* than the protein found in terrestrial vertebrates! There are two lessons here, the more important of which is that we must always remember that in any divergence there are opportunities for change along both lines of descent. Clearly the lamprey albumin has experienced a number of further internal duplications since cyclostomes and other vertebrates diverged. The second lesson is a reminder of another principle in protein evolution: Duplication begets more duplication. This has to do with the increased opportunities for DNA mis-matching between similar sequences. These in turn can lead to more unequal crossing over, for example (for a fuller discussion, see Doolittle, 1979). We are currently trying to clone the lamprey albumin in order to obtain its complete sequence. With that we should be able to pinpoint precisely when the first duplications took place, events that must pre-date the divergence of cyclostomes and other vertebrates. Like fibrinogen, albumin molecules ought to exist among the protochordates and some invertebrates.

APOLIPOPROTEINS

During our studies on lamprey albumin, we uncovered two small polypeptide contaminants that turned out to be high density apolipoproteins. In fact, these two proteins are among the most abundant proteins in lamprey plasma (Fig. 5). We were able to clone both of them from a lamprey liver cDNA library. Their complete sequences were determined, the larger, composed of 168 amino acids, by Manuel Pontes, a graduate student in our laboratory, and the smaller one, amounting to 76 residues, by Dr. Xun Xu, a visitor from China (Pontes *et al.*, 1987). These two proteins have many of the attributes of the high density apolipoproteins found in mammalian plasma, including structures that are highly helix-permissive and that lack cysteine. Although we were able to align the sequences with some mammalian apolipoprot-

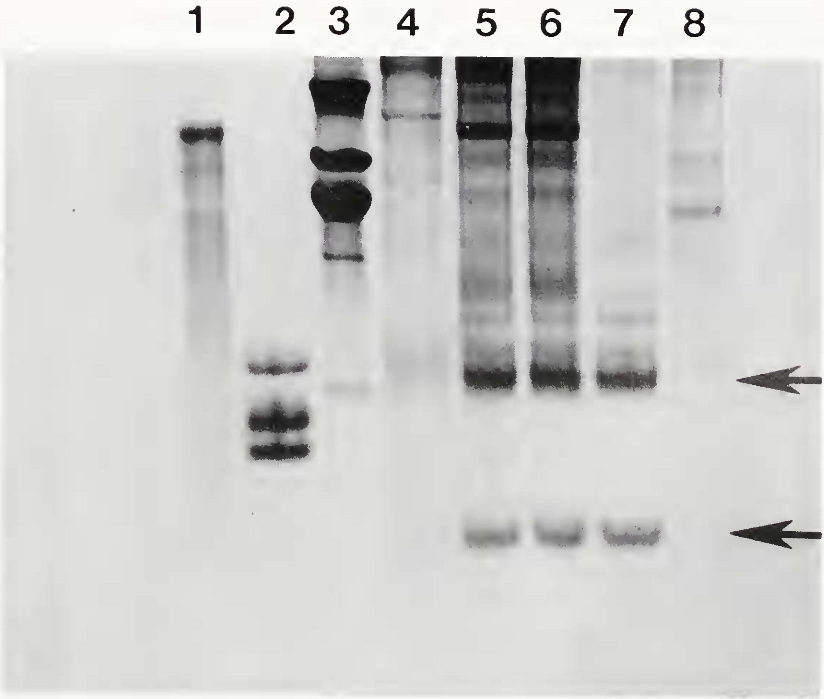

FIGURE 5. SDS-polyacrylamide (7.5%) gel electrophoresis (reducing conditions) of lamprey plasma and plasma materials. 1, purified lamprey transferrin; 2, reference substances (chymotrypsinogen, myoglobin and lysozyme); 3, lamprey fibrinogen; 4, lamprey albumin; 5 and 6, lamprey plasma; 7, high density ultracentrifugate (HDL layer); 8, lipoprotein "cake" from ultracentrifuged lamprey plasma (from Pontes *et al.*, 1987).

eins, the high rate of sequence change in these materials makes it difficult to prove a case for common ancestry.

TRANSFERRIN

Transferrin, the iron transport protein, is well known to be a major component in the blood plasmas of all vertebrates. It is also common knowledge that among all classes of vertebrates, including cyclostomes, the molecule is an internal dimer with two iron-binding sites. One of our graduate students, Barbara Evans, purified lamprey transferrin, and, together with Dr. Kenneth Watt, determined the amino-terminal 47 amino acids (Evans *et al.*, 1984). Attempts to clone the message were unsuccessful, however. We had hoped to shed more light on when the tandem duplication occurred that led to the double-sized molecule. In the meantime, an iron-binding protein has been found in the ascidian *Pyura stolonifera* that is half the size of vertebrate transferrin (Huebers *et al.*, 1984). It will be of great interest to see if the sequence of the ascidian protein is recognizably homologous with the vertebrate molecule.

COMPUTER SEARCHING

Up to this point, I have been discussing the evolution of vertebrate plasma proteins on the basis of comparisons of proteins as they exist in contemporary mammals on the one hand, and in lampreys, the most primitive of the vertebrates, on the other. I'd like now to turn to another approach, one in which the principal tool is the computer. The basic idea is to look for relationships between and among proteins simply on the basis of their amino acid sequences. Let me start with one of my favorite examples. A few years ago, the amino acid sequence of rat angiotensinogen was published. Angiotensinogen is the precursor of the hormone angiotensin, a 10-residue peptide that is critical for the maintenance of water balance. Interestingly, the precursor of this short peptide is enormous, amounting to no less than 453 amino acids in length. Nevertheless, I typed the entire sequence into my computer and searched it against a data base I was maintaining. Unexpectedly, the computer reported that angiotensinogen is related to α-1-antitrypsin (Doolittle, 1983). Overall, the two proteins were only a little more than 20% identical, but the fact that the resemblance extended over the course of almost 400 residues made the match-up quite significant. Now, α-1-antitrypsin is a protease inhibitor that is one of the abundant proteins in blood plasma (Table I), and the fact that it could share common ancestry with a polypeptide hormone precursor was astonishing to me. It was already known that α-1-antitrypsin is related to several other protease inhibitors, and since that time more inhibitor members of the family, which are now referred to as "serpins," have been identified. But the utilization of a large, apparently derelict, protease-inhibitor as the precursor of a tiny polypeptide hormone remains a unique example of the re-utilization of good stable proteins in new settings with quite different functional demands.

Since that time many other unexpected match-ups have resulted from the computer-searching of new sequences through data banks, and many of these have involved the vertebrate plasma proteins. The blood coagulation Factors VIII and V, for example, which are themselves homologous, were found to share common ancestry with the copper-binding protein ceruloplasmin (Vehar *et al.*, 1984). Less unexpectedly, for those following the structure and functional relationships of the plasma proteins, was the report that α-2-macroglobulin is homologous with the complement components C3 and C4 (Sottrup-Jensen *et al.*, 1985). Indeed, I myself had previously predicted this relationship on the basis of partial sequences and other features these molecules have in common (Fig. 6).

The sequence of α-2-macroglobulin, which appears to be the root-ancestor of the group, is interesting on other grounds. It is a very long sequence, consisting of over 1450 amino acid residues, and yet there are no residual signs of past internal duplications. The ordinary way for proteins to become elongated, as we have noted previously for albumin and transferrin, is by internal duplication. The absence of any vestige of internal duplication in α-2-macroglobulin implies that it is either very old or is changing very fast, or both. As it happens, yesterday afternoon I met Jim Quigley out in front of the MBL, and he told me that he and Peter Armstrong have purified a homologue of α-2-macroglobulin from the hemolymph of the horseshoe crab (Quigley and Armstrong, 1985). It will be very interesting to see what the sequence of the *Limulus* protein is like compared with the mammalian counterparts. At the very least the matter of its rate of change should be settled.

Unanticipated sequence resemblances among the plasma proteins continue to mount. The vitamin D-binding protein known as the "group-specific component" has recently been found to be related to albumin, the sequences being 24% identical (Yang *et al.*, 1985). Further, some plasma proteins whose functions had not previously been known, are now being classified on the basis of their relationship to other proteins. A minor component known as the "gamma-trace protein" is clearly related to the kininogen family, and kininogens in turn are now known to be related to thiol protease-inhibitors (Doolittle, 1985b; Muller-Esterl *et al.*, 1985).

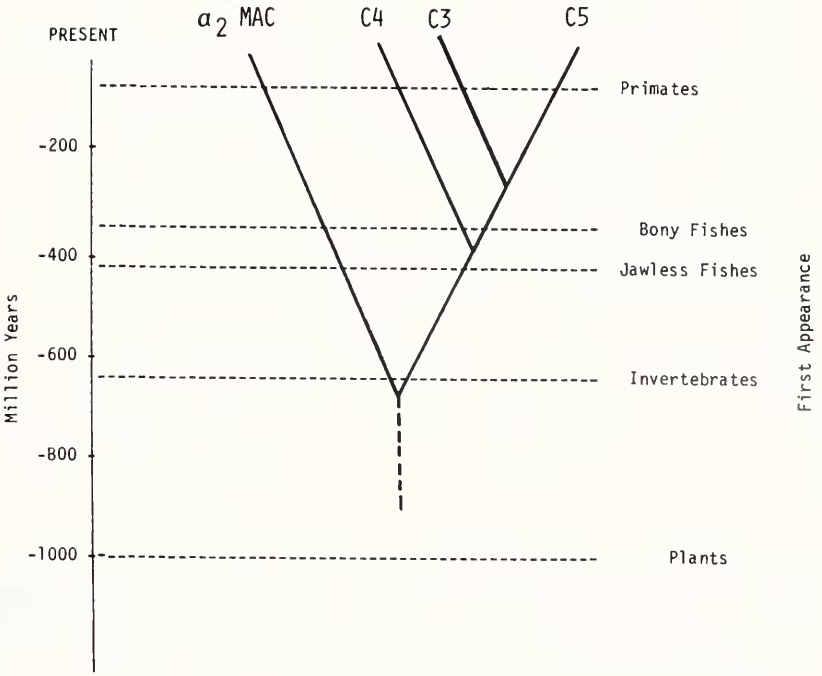

FIGURE 6. Possible scheme of gene duplications leading to present-day α-2-macroglobulin and three complement proteins. There are fewer species comparison data available for this tree than there are for the hemoglobin and fibrinogen trees in Figures 1 and 2, and, as a result, the divergence times are not nearly as accurate (from Doolittle, 1984)

EXON SHUFFLING

The resemblances that are being uncovered do not always extend over the full lengths of two similarly sized proteins. Sometimes, in fact, two proteins will have obviously similar segments over a portion of their lengths, and then, abruptly, the similarity is lost. In some cases one of the sequences may even switch to looking like a part of some third protein. One such case is observed in the protein known as tissue plasminogen activator (TPA). This protease precursor has short segments near its amino terminus that resemble, successively, fibronectin, epidermal growth factor, and prothrombin (Banyai *et al.,* 1983).

At about the same time the curious mosaic form of TPA was noticed, the full sequence of the epidermal growth factor precursor was reported. When we searched this very long sequence (more than 1200 amino acid residues) against our sequence collection, we were surprised to find that the candidate sequences retrieved were all blood clotting factors: Factor IX, Factor X, and Protein C (Doolittle *et al.,* 1984). The resemblances were limited to two 40–45 residue segments in the clotting proteins, but 10 similar segments appeared in the EGF-precursor (Fig. 7). Curiously, four of the EGF-precursor segments were more closely related to one of the segments in the clotting factors, while the six others, including EGF itself, were more closely related to the second. This isn't what one expects from a simple homologous crossing over; rather, it suggests repeated exchanges (Doolittle *et al.,* 1984).

Shortly thereafter, the laboratory of Brown and Goldstein reported the sequence

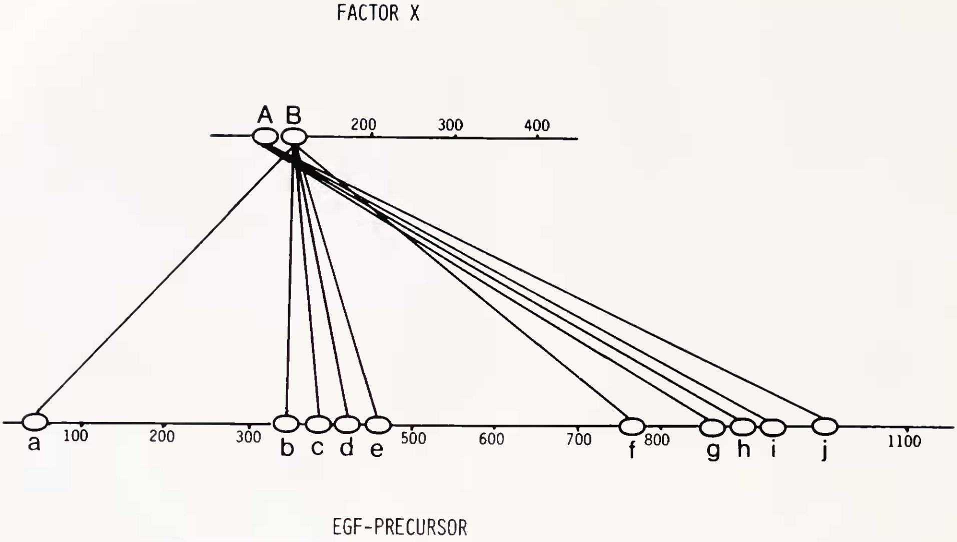

FIGURE 7. Schematic depiction of resemblances between segments found in the blood clotting protein Factor X and the membrane-associated protein which is the epidermal growth factor (EGF) precursor. Each of the oval segments represents a polypeptide segment of 40–45 residues, most of which contain six cysteines. EGF itself is represented by the right-most oval in EGFP.

of the LDL-receptor (Sudhoff *et al.*, 1985). Astonishingly, this receptor protein had a portion of about 200 amino acids that were strikingly similar to a segment of the EGF-precursor (Fig. 8). Moreover, another long region had seven repeated segments, the sequences of which were all similar to two segments found in the complement protein C9 (Fig. 8). Obviously DNA was being swapped around in a way that was putting similar peptide segments in various proteins (Table V).

The major clue for explaining how these exchanges must be occurring originally came from Ny *et al.* (1984) who determined the genomic DNA sequence corresponding to TPA. What they found was that the segments corresponding to other protein types, as identified by Banyai *et al.* (1983)—fibronectin, EGF, and prothrombin—were separated in each case by introns. Subsequently other workers found virtually the same pattern in the genes of the EGF-precursor and the LDL-receptor. In other words, each of the symbols corresponding to one of the prototype segments in Figure 8 actually corresponds to an exon. In these cases, exons must actually correspond to independently folding domains that can be shuffled about from protein to protein without disrupting the rest of the protein structure.

So far, about a half dozen such exchangable modules have been identified. They range in size from 40–80 amino acid residues. The most recent one to be characterized has already been found in several plasma proteins, including β-2-glycoprotein, complement factor B and complement factor H (Ripoche *et al.*, 1986). Moreover, at a recent Cold Spring Harbor meeting I learned from Earl Davie that the Factor XIII b-chain also contains a long series of segments of this type (E. Davie, pers. comm.).

OVERVIEW

It appears that the proliferation of vertebrate plasma proteins has been due to two related but distinguishably different phenomena. The first is orthodox gene duplica-

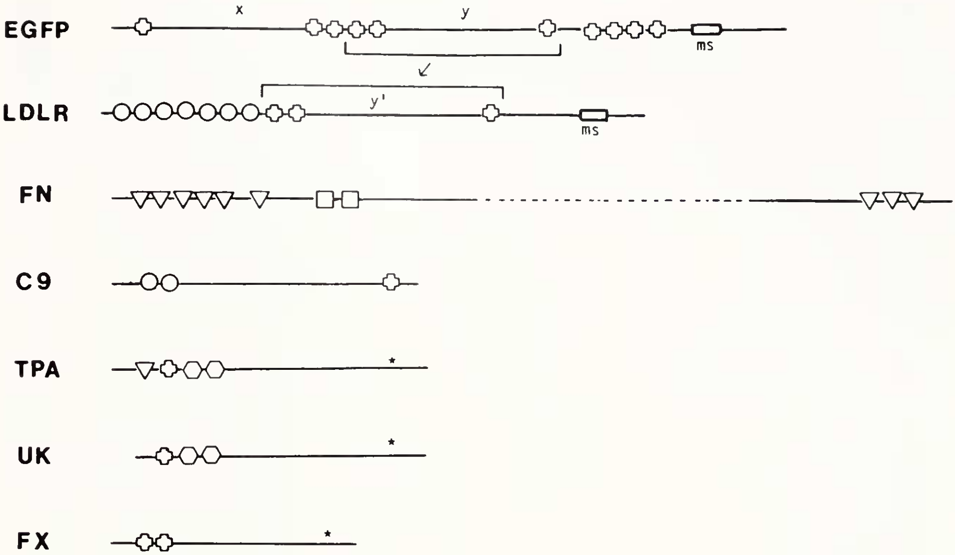

FIGURE 8. Schematic comparison of several vertebrate proteins that contain homologous "modules" resulting from exon shuffling. EGFP, epidermal growth factor precursor; LDLR, low-density lipoprotein receptor; FN, fibronectin; C9, complement component 9; TPA, tissue plasminogen activator; UK, urokinase; FX, blood clotting Factor X. The different types of peptide segments are denoted by the different symbols, asterisks (*) are the active sites of serine proteases, and "ms" marks membrane-spanning segments (from Doolittle, 1985a).

tion, complete versions of which can give rise to new, modified gene products, while partial or incomplete versions give rise to elongated, internally duplicated sequences. The second proliferative force is exon shuffling, an operation that has led to a relatively small set of polypeptide segments being present in many different proteins (Table IV). Most of these segments likely act as recognition units for binding the proteins

TABLE IV

Some animal proteins that have sequence segments in common

A. *EGF-type:*
Epidermal growth factor precursor
Tumor growth factors
Low density lipoprotein receptor
Factor IX
Factor X
Protein C
Tissue plasminogen activator
Urokinase
Complement C9
Notch protein (Drosophila)
Lin-12 (Nematode)
Thrombospondin

B. *Fibronectin "finger":*
Fibronectin
Tissue plasminogen activator

C. *C9-type:*
Complement C9
Low density lipoprotein receptor
(Fibronectin?)
Notch (Drosophila)
Lin-12 (Nematode)

D. *Proprotease "kringle":*
Plasminogen
Tissue plasminogen activator
Urokinase
Prothrombin

E. *β-2 type:*
β_2-glycoprotein
Complement Factor B
Complement Factor H
Factor XIII b-chain

TABLE V

Some known protein families found in vertebrate blood plasma

A. Albumin
   α-fetoprotein (fetal albumin)
   Vitamin-D-binding protein (group-specific component)

B. Immunoglobulins (assorted)
   β_2-microglobulin

C. Fibrinogen α chain
   Fibrinogen β chain
   Fibrinogen γ chain

D. Coagulation serine proteases
   Fibrinolysis serine proteases
   Complement serine proteases
   Miscellaneous serine proteases
   Haptoglobin

E. α_2-macroglobulin
   Pregnancy zone protein
   Complement C3
   Complement C4
   Complement C5

F. β_2-Glycoprotein
   Complement Factor B (nonenzyme portion)
   Complement Factor H
   Factor XIII b-chain

G. Retinol-binding protein
   α_1-microglobulin

H. Antithrombin III
   α_1-antitrypsin
   α_1-antichymotrypsin
   Angiotensinogen

I. Ceruloplasmin
   Factor V
   Factor VIII

J. Lipid-binding proteins
   (A, B, C, etc.)

K. Kininogens
   β_1-Microglobulin
   Acute phase proteins
   Thiol protease inhibitors

L. Transthyretin (prealbumin)
   Glucagon
   Glycentin

M. β-Thromboglobulin
   Platelet factor 4

N. Serum amyloid P-component
   C-reactive component

differentially to various cells. In any case, the net result is that the six hundred electrophoresis components observed in mammalian plasma will ultimately be grouped into a relatively small number of families (Table V).

When did all these events transpire? Doubtless they occurred throughout the history of the vertebrates, but the fact is that much of the inventive action must have been a necessary precondition to the evolution of vertebrates themselves. Many of the sequence comparisons we have discussed this evening indicate that the original gene duplication leading to some of the principal plasma proteins must have occurred among invertebrate ancestors. The point is made further by the fact that the lamprey, one of our most distant vertebrate relatives, has so many plasma proteins in common with mammals. The fact that many of these proteins have not yet been identified among invertebrates or protochordates remains a mystery, but it also provides an opportunity for much more exploration. I have no doubt that that matter will resolve itself as more sequence data are collected, particularly among the protochordates and echinoderms.

ACKNOWLEDGMENTS

I am grateful to all the many students and associates who have worked with me on these problems over the past 25 years. I thank Karen Anderson for assistance in preparing this manuscript.

LITERATURE CITED

ANDERSON, N. L., R. P. TRACY, AND N. G. ANDERSON. 1984. High resolution two-dimensional electrophoretic mapping of plasma proteins. Pp. 221–270 in *The Plasma Proteins,* Second ed, Vol. IV, F. W. Putnam, ed. Academic Press, New York.

Banyai, L., A. Varadi, and L. Patthy. 1983. Common evolutionary origin of the fibrin-binding structures of fibronectin and tissue-type plasminogen activator. *FEBS Lett.* **163**:37–41.

Bohonus, V. L., R. F. Doolittle, M. Pontes, and D. D. Strong. 1986. Complementary DNA sequence of lamprey fibrinogen beta chain. *Biochemistry* **25:** 6512–6516.

Brown, J. R. 1976. Structural origins of mammalian albumin. *Fed. Proc.* **35:** 2141–2144.

Doolittle, R. F. 1979. Protein evolution Pp. 1–118 in *The Proteins,* 2nd ed., Vol. IV, H. Neurath and R. L. Hill, eds. Academic Press, New York.

Doolittle, R. F. 1983. Angiotensinogen is related to the antitrypsin-antithrombin-ovalbumin family. *Science* **222:** 417–419.

Doolittle, R. F. 1984. Evolution of the vertebrate plasma proteins. Pp. 317–360 in *The Plasma Proteins,* Second ed., Vol. IV, F. W. Putnam, ed. Academic Press, New York.

Doolittle, R. F. 1985a. The genealogy of some recently evolved vertebrate proteins. *Trends Biochem. Sci.* **10:** 233–237.

Doolittle, R. F. 1985b. More homologies among the vertebrate plasma proteins. *Biosci. Rep.* **5:** 877–884.

Doolittle, R. F., J. L. Oncley, and D. M. Surgenor. 1962. Species differences in the interaction of thrombin and fibrinogen. *J. Biol. Chem.* **237:** 3123–3127.

Doolittle, R. F., and G. M. Fuller. 1972. Sodium dodecyl sulfate polyacrylamide gel electrophoresis studies on lobster fibrinogen and fibrin. *Biochim. Biophys. Acta* **263:** 805–809.

Doolittle, R. F., D. F. Feng, and M. S. Johnson. 1984. Computer-based characterization of epidermal growth factor precursor. *Nature* **307:** 558–560.

Evans, B. R., K. W. K. Watt, and R. F. Doolittle. 1985. Characterization and amino terminal sequence of lamprey transferrin. *Fed. Proc.* **69:** 3799 (abstr).

Fuller, G. M., and R. F. Doolittle. 1971. Transformation of lobster fibrinogen into fibrin. *Biochemistry* **10:** 1311–1315.

Huebers, H. A., A. W. Martin, and C. A. Finch. 1984. A mono-sited transferrin from a representative deuterostome: The ascidian *Pyura stolonifera. Fed. Proc.* **43:** 2058 (abstr).

Kuyas, C., M. Riley, J. Bubis, and R. F. Doolittle. 1983. Lamprey plasma albumin is a glycoprotein with a molecular weight of 175,000. *Fed. Proc.* **42:** 2085 (abstr).

Muller-Esterl, W., H. Fritz, J. Kellerman, F. Lottspeich, W. Machleidt, and V. Turk. 1985. Genealogy of mammalian cysteine proteinase inhibitors. Common evolutionary origin of stefins, cystatins and kininogens. *FEBS Lett.* **191:** 221–226.

Ny, T., F. Elgh, and F. Lund. 1984. The structure of the human tissue-type plasminogen activator gene: Correlation of intron and exon structures to functional and structural domains. *Proc. Natl. Acad. Sci. USA* **81:** 5355–5359.

Pontes, M., X. Xu, D. Graham, M. Riley, and R. F. Doolittle. 1987. cDNA sequences of two apolipoproteins from the lamprey. *Biochemistry* **26:** 1611–1617.

Quigley, J. P., and P. B. Armstrong. 1985. A homologue of α-2-macroglobulin purified from the hemolymph of the horseshoe crab *Limulus polyphemus. J. Biol. Chem.* **260:** 12715–12719.

Ripoche, J., A. J. Day, A. C. Willis, K. T. Belt, R. D. Campbell, and R. B. Sim. 1986. Partial characterization of human complement factor H by protein and cDNA sequencing: homology with other complement and non-complement proteins. *Biosci. Rep.* **6:** 65–72.

Sottrup-Jensen, L., T. M. Stepanik, T. Kristensen, P. B. Lonblad, C. M. Jones, D. M. Wierzbicki, S. Magnusson, H. Domdey, R. A. Wetsel, A. Lundwall, B. F. Tack, and G. H. Fey. 1985. Common evolutionary origin of α-2-macroglobulin and complement components C3 and C4. *Proc. Natl. Acad. Sci. USA* **82:** 9–13.

Strong, D. D., M. Moore, B. A. Cottrell, V. L. Bohonus, M. Pontes, B. Evans, M. Riley, and R. F. Doolittle. 1985. Lamprey fibrinogen gamma chain: Cloning, cDNA sequencing, and general characterization. *Biochemistry* **24:** 92–101.

Sudhoff, T. C., D. W. Russell, J. L. Goldstein, M. S. Brown, R. Sanchez-Pescador, and G. I. Bell. 1985. Cassette of eight exons shared by genes for LDL receptor and EGF precursor. *Science* **228:** 893–895.

Vehar, G. A., B. Keyt, D. Eaton, H. Rodriguez, D. P. O'Brien, F. Rotblat, H. Oppermann, R. Keck, W. I. Wood, R. N. Harkins, E. G. D. Tuddenham, R. M. Lawn, and D. J. Capon. 1984. Structure of human factor VIII. *Nature* **312:** 337–342.

Yang, F., J. L. Brune, S. L. Naylor, R. L. Cupples, K. H. Naberhaus, and B. H. Bowman. 1985. Human group-specific component (Gc) is a member of the albumin family. *Proc. Natl. Acad. Sci. USA* **82:** 7994–7998.